Design and testing for clinical trials faced with misclassified causes-of-death

BART VAN ROMPAYE*,[1], SHABBAR JAFFAR[2], ELS GOETGHEBEUR[1]

[1] Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, S9, 9000 Ghent, Belgium

[2]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.

Abstract With clinical trials under pressure to produce more convincing results faster, we re-examine relative efficiencies for the semi-parametric comparison of cause-specific rather than all-cause mortality events, observing that in many settings misclassification of cause-of-failure is not negligible. By incorporating known misclassification rates, we derive an adapted logrank test which optimizes power when the alternative treatment effect is confined to the cause-specific hazard. We derive sample size calculations for this test as well as for the corresponding all-cause mortality and naive cause-specific logrank test which ignores the misclassification. This may lead to new options at the design stage which we discuss. We re-examine a recently closed vaccine trial in this light and find the sample size needed for the new test to be 32% smaller than for the equivalent all-cause analysis, leading to a reduction of 41,224 participants.

Cause-specific analysis; Clinical trials; Competing risks; Misclassification; Sample size; Survival analysis; Verbal autopsy.

1 Introduction

Clinical trials with survival outcomes typically study products designed to reduce the hazard of some targeted cause-specific event. For instance, targeted cancer therapies aim to reduce the hazard of cancer-specific mortality without affecting mortality from other causes (Cuzick, 2008). Vaccine trials hope to reduce specific morbidity, but are not expected to prevent all types of disease. In such case it is well known that a logrank comparison

[1]* To whom correspondence should be addressed: bart.vanrompaye@ugent.be, Tel : + 32 9 264 47 56, Fax : + 32 9 264 49 95

of all-cause mortality involves a diluted version of the cause-specific hazard ratio and therefore loses power. The logrank test limited to cause-specific events is then generally the preferred approach. In practice, the trade-off between both tests needs to address feasibility and cost of cause assessment as well as relative efficiency. This balancing act is part of standard design considerations in any clinical trial with survival outcomes.

While efforts have been made to accommodate a proportion of missing failure types in this setting (e.g. Rowe, 2006; Goetghebeur and Ryan, 1990; Lu and Tsiatis, 2005) misclassification of causes-of-event is rarely acknowledged. This happens frequently however, especially with verbal autopsies used in developing countries (Soleman, Chandramohan and Shibuya, 2006) and for events which are generally difficult or costly to diagnose (e.g. onset of Alzheimer's disease, Waldemar *and others* 2007), but also when using death registries or administrative databases in the study of common causes of death in Western countries (e.g. Ladouceur *and others*, 2007). Uncertainty on the cause-of-death may drain the power of a 'naive' cause-specific analysis based on observed causes-of-death and confound estimators (as shown in Anker, 1991; Maude and Ross, 1997). The absence of such effects in an all-cause analysis may shift the balance of preference between both types of analysis. Jaffar *and others* (2003) describe a vaccine trial where frequent misclassification could lead to such severe loss of power that the primary endpoint would better be changed from acute lower respiratory tract infection (ALRI)-specific mortality to all-cause mortality.

In this paper, we demonstrate how anticipated misclassification rates can be incorporated in the analysis, thus deriving an adapted cause-specific logrank test which recovers some of the power loss. To this end we model proportional cause-specific hazards alternatives with possibly misclassified observed failure patterns. A general test statistic is then derived from a par-

tial likelihood involving weighted contributions from all observations. Under simplifying assumptions this becomes an intuitive expression with weights according to the failure type, an approach reminiscent of the one by Goetghebeur and Ryan (1990) for missing causes-of-death. This provides an efficient alternative test to the current options for the intention-to-treat analysis. We revisit the ALRI vaccine trial from this perspective and study how much lower the needed sample size could have been for the adapted test.

In section 2 we motivate the problem in more detail. In section 3 we define notation and the alternative hypothesis for which we seek better power. We derive the adapted logrank test in section 4, sample size implications are considered in section 5. Finally, we apply the new test statistic to data from the Gambia Pneumococcal Vaccine Trial in section 6. Issues of sensitivity and deviations from model assumptions are considered in section 7. Selected details can be found in the supplementary materials (http://www.biostatistics.oxfordjournals.org).

2 The Gambia Pneumococcal Vaccine Trial

Yearly more than one million children under the age of 5 die of acute respiratory infections caused by pneumococci (WHO, 1999). A large-scale randomized, double blind trial (the Gambia Pneumococcal Vaccine Trial) evaluated the effectiveness of a pneumococcal conjugate vaccine in the developing country setting of eastern parts of The Gambia, where the rate of childhood ALRI is up to tenfold higher than in industrialized countries (O'Dempsey *and others*, 1996). Jaffar *and others* (1997) describe the mortality patterns in this region. Final study results are published in Cutts *and others* (2005).

Initially the study focused on ALRI mortality, with cause-of-death (COD) generally not determined clinically but through 'verbal autopsy' or 'post-mortem questionnaires'. Here, a team of three doctors assigns COD from

data on the sequence and duration of the signs and symptoms preceding death, gathered by retrospectively interviewing the deceased's caretakers (De Francisco *and others*, 1993). Even with two out of three doctors agreeing on COD this method has low sensitivity for most CODs (sometimes below 50%, e.g. Snow *and others* 1992 on malaria and ALRI) and misclassification is common. Prevalence of misclassification when using verbal autopsy is reviewed in Anker (1991) or Maude and Ross (1997). An extensive review of methods for verbal autopsy is found in Soleman *and others* (2006), while Chandramohan *and others* (2005) formulate general concerns. In our work we start from known misclassification probabilities.

Since the vaccine directly targets pneumococci little prevention of deaths from causes other than ALRI is expected, even though the vaccine likely prevents deaths from invasive pneumococcal diseases such as meningitis and bacteraemia. Combined with high misclassification rates this could substantially dilute the estimated treatment effect in a naive cause-specific analysis, which thus loses power. As in Snow *and others* (1992), Jaffar *and others* (2003) assume a sensitivity of 40% and specificity of 90% and show this decreases the power in the Gambian setting from an expected 93% in the absence of misclassification to an expected 54%. The initial plan to examine the vaccine impact on ALRI mortality thus changed focus to all-cause mortality. However, the larger sample size requirement due to a diluted effect, combined with the ethical desire to speed up trial completion, caused another change in endpoint to disease-free survival with radiologically confirmed pneumonia as endpoint (for results see Cutts *and others* 2005). Since misclassification steered the primary endpoint away from a cause-specific interpretation, correcting for it might bring back a viable cause-specific analysis.

Correcting for misclassification has happened for cause-specific mortality fractions (Anker, 1991) and cause-specific mortality rates (Maude and Ross,

1997). Archer and Ryan (1989) correct for misclassification in the cause-of-death test for carcinogenicity using a missing data approach. Ebrahimi (1996) adopted a fully parametric Bayesian approach to fit competing risk models. Finally, Dewanji and Sengupta (2003) developed an EM-algorithm to estimate cause-specific hazards non-parametrically when information is missing at random but give no simple test of treatment effects. The MAR-assumption does not allow for misclassification depending on the true cause-of-death. They also introduce a Nelson-Aalen type estimator assuming one always partially recognizes the true cause-of-death in the diagnosis, a setting different from ours.

3 Notation and model assumptions

We consider the cause k-specific hazard for an individual $i \in \{1, \ldots, n\}$

$$h_k(t; Z_i) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(D_i < t + \Delta t, \delta_i = k | D_i \geq t; Z_i)$$

where D_i is the time to failure, $\delta_i = k$ is the true failure type which is 0 (other causes) or 1 (cause of interest) and Z_i is a binary covariate, typically a treatment: 0 for control and 1 for treatment with randomization probability $P(Z_i = 1) = \pi$. The null hypothesis of interest is $h_1(t; 1) = h_1(t; 0)$, i.e. no treatment effect on the type 1-specific hazard. Due to misclassification, as indicated by M_i (1 if the failure type is misclassified and 0 otherwise), these hazards cannot be observed directly. The observed failure type in the absence of censoring is $F_i = (\delta_i - M_i)^2$.

We assume non-informative censoring can occur (meaning net and crude cause-specific hazards coincide, Fleming and Harrington, 1991) in which case $C_i = 0$ (and 1 otherwise). The observation time is denoted T_i, and for any individual i we observe T_i, C_i, and Z_i, and if $C_i = 1$ also F_i.

As for missing data, some untestable assumption about the nature of the errors is inevitable. We assume the misclassification probabilities are known and may depend on failure time D_i, true failure type δ_i and treatment Z_i.

Assumption 1:

$$pr(M_i = 1 | D_i = t_i, C_i = 1, \delta_i = k, Z_i = z_i) = p_k(t_i; z_i)$$

Such dependence on the true, unobserved failure type is often realistic. Further expressions simplify substantially if these probabilities (then denoted $p_k(t)$) do not depend on Z_i and we develop our results in those terms without loss of generality. We call $1 - p_1(t)$ the sensitivity of the cause-of-death diagnosis and $1 - p_0(t)$ the specificity.

We will seek to gain power targeting an alternative where Z has no influence on the type 0-specific hazard but has a proportional effect on the disease-specific hazard.

Assumption 2:

$$h_0(t; Z) = h_0(t)$$
$$h_1(t; Z) = e^{\phi Z} h_1(t)$$

with $h_k(t) = h_k(t; 0)$ the baseline hazard for a type k failure. While the second equation is a standard proportional hazards assumption, the first equation presents a stronger assumption that can be relaxed using a Cox-type model, which is beyond the scope of this paper. Our method will have optimal power under these assumptions, but is still valid for different true alternatives.

Finally, we formalize a connection between the two cause-specific baseline hazards.

Assumption 3:

$$h_1(t) = e^{\xi(t)} h_0(t)$$

As long as $e^{\xi(t)}$ is arbitrary no restrictions are imposed on the different hazards. However, for simplicity we will later choose a parametric shape, the simplest being a constant ratio of risks attributable to different failure types.

Assumption 3':

$$h_1(t) = e^{\xi} h_0(t)$$

More flexible parametric forms are discussed in section 4 of the supplementary materials (http://www.biostatistics.oxfordjournals.org).

4 Derivation of the test statistic

Under assumptions 1-3, a partial likelihood is based on the conditional probabilities of observing one of two event types at time t_i: $F_i = 0$ or 1, given one such event is observed in the risk set at t_i. We first assume $\xi(t)$ is known, for example from a pilot study. With $\mathcal{R}_i$ the set of subjects who had not failed or been censored just prior to failure time t_i, the partial likelihood becomes:

$$\begin{aligned} L_p\,(\text{observed data}) \;=\; & \prod_{\substack{i:c_i=1,\\ f_i=0}} \frac{h_1(t_i)\{e^{-\xi(t_i)}(1-p_0(t_i)) + e^{\phi z_i} p_1(t_i)\}}{\sum_{j\in\mathcal{R}_i} h_1(t_i)\{e^{-\xi(t_i)}(1-p_0(t_i)) + e^{\phi z_j} p_1(t_i)\}} \\ & \prod_{\substack{i:c_i=1,\\ f_i=1}} \frac{h_1(t_i)\{e^{\phi z_i}(1-p_1(t_i)) + e^{-\xi(t_i)} p_0(t_i)\}}{\sum_{j\in\mathcal{R}_i} h_1(t_i)\{e^{\phi z_j}(1-p_1(t_i)) + e^{-\xi(t_i)} p_0(t_i)\}} \end{aligned}$$

The score statistic under the null then becomes:

$$T = \sum_{i:C_i=1} w_i\{t_i, F_i\}(Z_i - \bar{Z}_i) \tag{4.1}$$

with $\bar{Z}_i = \sum_{j\in\mathcal{R}_i} Z_j/n_i$ the mean of the covariate values for the n_i persons at risk at t_i and w_i a weight function depending on the event time t_i but also

on the observed failure type F_i:

$$w_i(t_i, F_i) = \begin{cases} \frac{1}{1+e^{-\xi(t_i)}\frac{1-p_0(t_i)}{p_1(t_i)}} & F_i = 0 \\ \frac{1}{1+e^{-\xi(t_i)}\frac{p_0(t_i)}{1-p_1(t_i)}} & F_i = 1 \end{cases} \tag{4.2}$$

The second weight equals $P(\delta_i = 1|F_i = 1)$, the positive predictive value of the diagnosis, the first weight is one minus the negative predictive value. Hence, type 1 observations are downweighted since they might actually be type 0 failures, while type 0 observations contribute to the statistic because they could really be type 1 failures.

Through the martingale central limit theorem the standardized test statistic $U = T/\left(V\right)^{1/2}$ can be shown to have an asymptotic standard normal distribution under the null, with

$$V = \sum_{i:C_i=1} w_i^2\{t_i, F_i\}\left[\left(\sum_{j\in\mathcal{R}_i} Z_j^2/n_i\right) - \bar{Z}_i^2\right] \tag{4.3}$$

Dewanji (1992) proposed another partial likelihood for missing causes-of-death, conditioning on any event type occurring (instead of the observed type). This leads to a minor gain in power at the expense of an increased complexity and a reduced robustness to model misspecification (Lu and Tsiatis 2005). We do not pursue this direction.

U depends on the relative cause-specific hazard $e^{-\xi(t)}$ through the weights w_i. If no a priori values are available for the $\xi(t_i)$ estimation is required, but if a consistent estimator is used in (4.2) the asymptotic distribution of U remains unchanged. While a nonparametric estimator based on a kernel-weighted partial likelihood can be derived, one can also obtain simple consistent estimators under parametric assumptions such as assumption 3' (both approaches are presented in the web-based supplementary material).

The remainder of this text assumes time-constant misclassification rates

p_0 and p_1 and uses assumption 3': $\xi(t) = \xi$. With O_k the total number of type k observations ($k = 0, 1$) the ξ-estimator becomes:

$$\widehat{e^{-\xi}} = \frac{O_1 p_1 - O_0(1 - p_1)}{O_0 p_0 - O_1(1 - p_0)} \tag{4.4}$$

This intuitive formula performs much better than the naive estimator O_0/O_1, reducing to it when $p_0 = p_1 = 0$.

Under the same assumptions T and V become weighted sums of classical logrank test contributions, enabling simple software implementation. For the numerators T_k and denominators $V_k^{1/2}$ of standard cause-specific logrank statistics with only COD=k as event, we have:

$$U = \frac{T}{(V)^{1/2}} = \frac{w_{F_i=0} T_0 + w_{F_i=1} T_1}{\left(w_{F_i=0}^2 V_0 + w_{F_i=1}^2 V_1\right)^{1/2}} \tag{4.5}$$

When p_0, p_1 and ξ are time-constant we denote the weights in (4.2) w_{F_i}. In the absence of misclassification $w_{F_i} = F_i$ and U simplifies to the classical cause-specific logrank statistic.

5 Sample size considerations

A main concern related to misclassification of the cause-of-death is how the loss of power for a cause-specific analysis leads to an increase in the needed sample size. Jaffar *and others* (2003) show for the pneumococcal vaccine trial how a cause-specific analysis can become less attractive than the corresponding all-cause analysis, prompting an (often undesirable) change of primary endpoint. Since our test statistic accounts for misclassification rates some power may be restored and the cause-specific analysis might be favoured from an efficiency perspective when specific alternatives are targeted.

This section illustrates the possible impact of misclassification on the

power and compares different possible tests at finite sample sizes by a simulation study and asymptotically. An appropriate sample size formula for the adapted test is derived.

5.1 Comparison of tests by simulation of the vaccine trial setting

To examine the loss of power at various misclassification rates we imitate the Gambian setting from Jaffar *and others* (2003), with 4 years of uniform accrual into the clinical trial and 0.5 years of additional follow-up. In the control group, the ALRI-specific death hazard $h_1 = 0.0059$ and the hazard of death from other causes $h_0 = 0.0275$ (derived from Jaffar *and others* 1997). In the treatment group the ALRI-specific hazard is reduced by 31.5% ($\phi = \log(0.685) = -0.378$). We further set π=0.5. For simplicity we use constant hazards and administrative censoring only. We vary p_1 from 0 to 60% and p_0 from 0 to 20%, both in steps of 2%, covering the 60% and 10% used by Jaffar *and others* (2003).

Sample size formula (5.1) is derived following Schoenfeld (1981) assuming equal randomization probabilities (as in Schulgen *and others* 2005 or Latouche and Porcher 2007). In the absence of misclassification $W = 1$ and we obtain a conservative sample size of 22,760 needed to get 80% power using a standard cause-specific logrank test at the 5% significance level.

Note that without treatment effect on the competing risk (assumption 2), all components of expression (4.1) have theoretical expectation 0 under the null, irrespective of the chosen weights. Hence, even under misspecification of p_0, p_1 and $\xi(t)$ the type I error rate is controlled. This was confirmed by simulation for the different tests considered (supplementary materials).

We consider three cause-specific logrank tests. The infeasible classical logrank test based on the true (in reality unobserved) failure type serves as

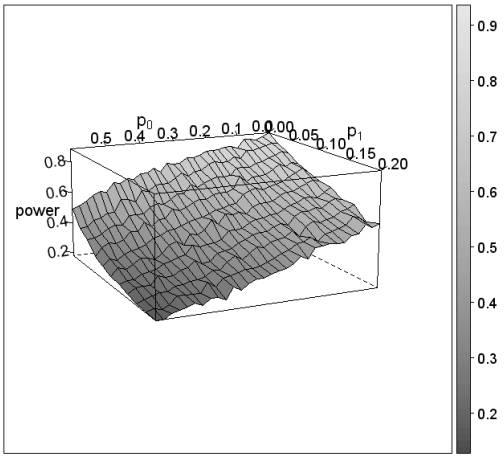


Figure 1: *Loss of power for the naive test as a function of the two misclassification rates* p_0 *and* p_1*.*

a reference and always yields a power of approximately 87% per design. For the 'naive' logrank test based on the observed (misclassified) failure types the power is expected to decrease substantially as the misclassification probabilities rise. The third test uses the adapted logrank statistic $U = T/(V)^{1/2}$ derived in (4.1) and (4.3), with ξ estimated through (4.4). These tests are called the reference, the naive and the adapted test, respectively.

Figure 1 gives the empirical power based on 1,000 simulations in function of the misclassification probabilities for the naive test. The standard error on the estimates is expected to be below 1%.

At the expected $p_0 = 10\%$ and $p_1 = 60\%$ the power drops from the reference 87% to only 25%, an extreme loss which can in part be attributed to the strong imbalance in mortality patterns ($\xi << 0$). The general features for the adapted test (not shown) are the same but with a higher power at all (p_0, p_1) combinations, leading to a power of 32% when $(p_0, p_1) = (10\%, 60\%)$. Although estimates for ξ from (4.4) can show extreme deviations under the alternative, using the true ξ in the adapted test leads to little or no difference in power. As table 1 will later show, the impact of using the adapted test statistic grows as the imbalance in mortality patterns is less pronounced.

The all-cause logrank test is unaffected by misclassification but has a power of merely 23% due to the diluted treatment effect. Thus, even at $p_0 = 10\%$ and $p_1 = 60\%$ the naive test has more power. Since the difference is small one may still prefer the all-cause analysis however, if the exact p_0 and p_1 are unknown. In contrast, the adapted cause-specific analysis is a more viable alternative since the difference with the all-cause logrank test is more pronounced. Section 5.2 confirms this feature analytically.

5.2 Asymptotic relative efficiencies

Assuming ξ, p_0 and p_1 to be constant over time a more direct comparison of the performance of the various tests is possible.

Consider the general weighted logrank test statistic

$$U = \frac{w^0 T_0 + w^1 T_1}{\{(w^0)^2 V_0 + (w^1)^2 V_1\}^{1/2}}$$

where super- and subscripts denote the observed failure type which determines the general weight factors w. Three special cases are:

1. the naive 'observed cause'-specific logrank statistic: $w^0 = 0$, $w^1 = 1$
2. the adapted logrank statistic: $w^0 = w_0$, $w^1 = w_1$
3. the all-cause logrank test: $w^0 = 1$, $w^1 = 1$

In the adapted statistic the change from super- to subscript indicates going from general weights to the weight factors w_{F_i} introduced at the end of section 4.

Under a sequence of contiguous alternatives

$$\lim_{n\to\infty} n^{1/2} \log\left(\frac{h_1(t;1)}{h_1(t)}\right) = \phi^* g(t)$$

Table 1: Asymptotic relative efficiencies for the three proposed tests.

	naive	adapted	all-cause
naive	1		
adapted	$1 + \frac{w_0 p_1}{w_1(1-p_1)}$	1	
all-cause	$\frac{1}{\left(1+e^{-\xi}\right)w_1(1-p_1)}$	$\frac{1}{\left(1+e^{-\xi}\right)(w_0 p_1 + w_1(1-p_1))}$	1

(where $g(t)$ is continuous on $[0, \tau]$ and the limit is achieved uniformly over $[0, \tau]$, the observation period) U is asymptotically normally distributed with unit variance and mean μ. Under the alternative $g(t) = 1$ (a constant treatment effect) the noncentrality parameter μ is:

$$\frac{\phi^* \left(w^0 p_1 + w^1(1-p_1)\right)}{\left[(w^0)^2 \left\{p_1 + (1-p_0)e^{-\xi}\right\} + (w^1)^2 \left\{(1-p_1) + p_0 e^{-\xi}\right\}\right]^{1/2}} \left[\int_0^\tau \frac{s_0(t)s_1(t)}{s(t)} h_1(t)dt\right]^{1/2}$$

as shown in the supplementary material (http://www.biostatistics.oxfordjournals.org). Here $s_i(t)$ is the limiting proportion of people at risk in treatment group i at time t over the total number in the study n, and $s(t) = s_1(t) + s_0(t)$.

By introducing the appropriate weights this leads to the asymptotic relative Pitman efficiencies $\text{ARE}(i, j) = \left(\frac{\mu_j}{\mu_i}\right)^2$, which represent the ratios of samples sizes asymptotically needed for two tests. The ARE's, shown in table 1, are determined by the relative strength of the competing causes and the severity of the misclassification. Figure 2 shows the ARE's as a function of p_0 and p_1 for $e^{-\xi} = 0.0059/0.0275$.

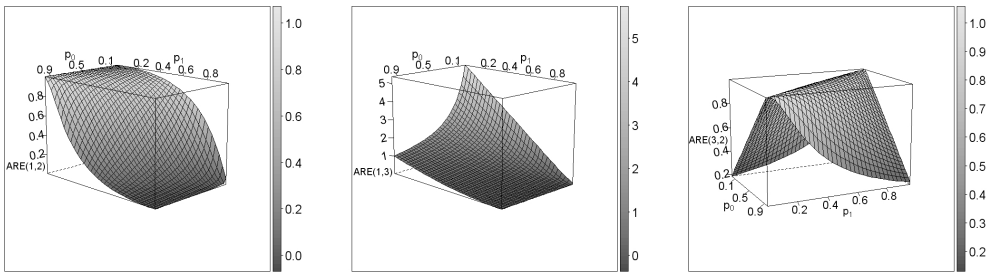
When misclassification is rare the naive test is more efficient than an all-cause analysis, but it loses efficiency quickly when p_0 and p_1 increase (figure 2(b)). This illustrates that a cause-specific analysis is preferred over a diluted all-cause analysis when the diagnosis is reliable.

The adapted cause-specific test is always at least as efficient as the naive one since w_0 and w_1 are both positive and $0 < p_1 < 1$. Only when the sen-

sitivity equals 1 the two tests are exactly the same ($w_0 = 0$). Figure 2(a) illustrates how with decreasing sensitivity the naive analysis loses efficiency much faster than the adapted test because the estimated treatment effect gets diluted, which is more pronounced for large p_0. The impact of using the adapted test depends on the mortality pattern through the ξ-dependence of the weights w_0 and w_1. For example, with p_0 small and p_1 large (as in section 5.1) the ARE becomes smaller with increasing $e^{-\xi}$. In summary, the adapted test always outperforms the naive one in the presence of misclassification, but this is modulated by various parameters in a complex way.

Finally, we are interested in the relative efficiency between the adapted and the all-cause analysis. The efficiency of the adapted test is always equal or higher to that of the all-cause analysis (figure 2(c)). Most is gained when p_0 and p_1 are both extreme in the same direction (both high or both low). At $p_0 = p_1 = 50\%$ the diagnosis is completely random and the two tests are identical, yielding an ARE of 1. In fact, when $p_0 + p_1 \approx 1$ the ARE approximates 1, an effect modulated by ξ: the more positive ξ, the more dominant the cause of interest, the smaller the dilution of the all-cause test and the wider the p_0, p_1-surface at which the ARE approximates 1.

To summarize, an adapted cause-specific analysis is always more efficient than an all-cause analysis, certainly making it an alternative to consider in the design of a study.



(a) ARE(1,2) (b) ARE(1,3) (c) ARE(3,2)

Figure 2: *Pitman ARE's between the three candidate statistics, at* $\xi =$ *log*$(h_1(t)/h_0(t)) =$ *log(0.0059/0.0275) as in Jaffar and others (2003).*

5.3 Sample size formula

The loss of power with misclassification affects sample size calculations for the cause-specific analysis. The noncentrality parameter for the adapted test statistic ($w^0 = w_0$ and $w^1 = w_1$) is:

$$\mu = \sqrt{n}\phi\sqrt{w_0 p_1 + w_1(1-p_1)}Q$$

where the integral expression Q involves the probability of seeing an event, which is approximated using the event rate under the alternative in the treated population. For constant hazards h_0 and h_1 a staggered accrual between time 0 and a and administrative censoring at time $a+f$ (as in Schulgen *and others* 2005 e.g.)

$$Q^2 = \frac{\pi(1-\pi)h_1 e^{\phi}}{h_1 e^{\phi} + h_0}\left(1 + \frac{e^{-(h_1 e^{\phi}+h_0)(a+f)} - e^{-(h_1 e^{\phi}+h_0)f}}{a(h_1 e^{\phi}+h_0)}\right)$$

The sample size formula for the adapted logrank test then becomes:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\phi^2 W^2 Q^2} \tag{5.1}$$

where $W = w_0 p_1 + w_1(1-p_1)$. For interventions which lower the hazard, this returns slightly conservative sample sizes recommended for design purposes. The difference with standard expressions lies in the factor W^2 which is 1 in the absence of misclassification. W thus allows to compare settings with and without misclassification.

5.4 Sample size and ARE for the Gambian illustration

We illustrate the use of sample size formula (5.1) by returning to the setting of Jaffar *and others* (2003): $h_1 = 0.0059$, $h_0 = 0.0275$, ϕ=log(0.685), p_0=10% and p_1=60%. Under our assumptions the naive analysis requires a

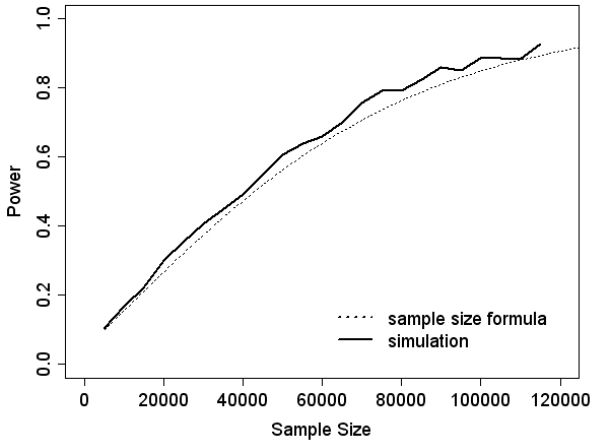


Figure 3: *Comparison between empirical and theoretical power.*

41% bigger sample size than our adapted analysis ($ARE(2,1)$=1.41). Since $ARE(3,1)$=0.96 the naive analysis needs just a slightly smaller sample than the all-cause analysis, so one may prefer the all-cause analysis when p_0 and p_1 may be misspecified. However, $ARE(3,2)$=0.68 meaning the all-cause test requires an approximately 50% larger sample than our adapted analysis, making the adapted test appealing even under mild misspecification of p_0 and p_1. This is illustrated by a detailed sensitivity analysis in section 5 of the supplementary material (http://www.biostatistics.oxfordjournals.org).

For the Gambian setting, figure 3 compares the empirical power for the adapted test with the power from formula (5.1), based on 1000 simulations at $\alpha = 5\%$.

At $\alpha = 5\%$ and $\beta = 20\%$ the needed sample size for the adapted test from (5.1) is 87,600. From the ARE this becomes 128,824 for the all-cause analysis, yielding an impressive absolute difference of 41,224. The sample size for the naive analysis is 123,516. Note that the decision to change the endpoint from mortality to morbidity was based on these huge sample sizes.

6 Application to data from the Gambia Pneumococcal Vaccine Trial

The various analyses were applied to data from the Gambia Pneumococcal Vaccine Trial (original results were published by Cutts *and others*, 2005). A sample of 17,433 individuals consisted of children born between september 1999 and the beginning of the year 2003, of which 8,715 got vaccine and 8,718 placebo. The vaccination usually took place between 40 and 400 days of age, with a median of 75 days. Follow-up stopped at the end of april 2004. Further structure in the data is ignored for the purpose of illustration.

The data contained information on 491 deaths in the control and 426 in the treatment group. Of these 917 deaths, 186 were classified as due to ALRI: 99 under control and 87 under treatment. The total follow-up time was 18,601 person years in the control group and 18,640 person years in the treatment group.

An all-cause logrank test for the treatment effect yields a p-value of 0.029, for a naive ALRI-specific logrank test this becomes 0.371. Assuming $p_0 = 10\%$ and $p_1 = 60\%$ the estimate for $e^{-\xi} = 1.92$, which differs from the value 4.66 derived from Jaffar *and others* (1997). Of course, for the purpose of testing our estimator was derived under the null which is rejected in this example. Nevertheless, we find the competing risks to be more prevalent. The adapted logrank statistic based on the estimated ξ is 1.921, yielding a borderline significant p-value of 0.055.

The adapted cause-specific analysis gained substantial power compared to a naive cause-specific analysis and yields a comparable (though slightly higher) p-value than the all-cause analysis, despite the higher ARE. Reduced power could come from model misspecification: Jaffar *and others* (2003) and Cutts *and others* (2005) suspect the vaccine may also influence the hazards

for some competing risks. This situation falls beyond the reach of our log-rank test and warrants further development into a Cox-type model, a topic of ongoing research. As it stands, our current correction still appears a competitive alternative for the all-cause comparison.

7 Discussion

This paper presents a test which corrects for known misclassification rates when comparing cause-specific survival between treatment groups and recovers power when treatment effect is confined to the disease-specific hazard. We argued how the added power may at the design stage tip the balance between choosing an all-cause versus cause-specific focus for the primary comparison. When considering the trade off, it is important to recognize the specificity of these two tests: the assumptions under which they typically operate and the types of errors that can be made in either case, under the null and under the alternative we anticipate here.

Standard group comparisons of all-cause hazards or of disease-specific hazards are not concerned with how the treatment effect is distributed over the different cause-specific hazards. For instance, equality between all-cause hazards, the standard null hypothesis of the all-cause analysis, need not imply strict equality between cause-specific hazards over the groups. Hence, rejection of the cause-specific null need not imply rejection of the all-cause null hypothesis and it could be argued that the cause-specific test is not valid for an all-cause comparison since it compromises its type I error. While technically correct, contrasting treatment effects on both competing hazards would need to cancel each other out exactly at every time point in order to produce identical all-cause hazards over time. The possibility of such highly unlikely and unstable equilibrium point is easily ignored assuming faithfulness as in causal inference (Spirtes *and others*, 1999).

In the opposite direction, there is the risk of missing an effect on all-cause mortality when performing a cause-specific analysis because the true effect is not confined to the cause of interest as anticipated. As long as a study is powered adequately to detect the partial effect on the cause of interest, it will enjoy the benefit of added power in the presence of a synergistic effect on the competing risk. In general, since design assumptions are not guaranteed to hold, ethical considerations invariably warrant a complementing all-cause analysis and verification of whether the observed effect on the cause of interest is dampened, reversed or indeed emphasized by the effect on the competing hazard. Recognizing the distinct effects on competing risks in the presence of misclassified causes of death requires adapted cause-specific (Cox) regression modelling which is the topic of further research.

Without claiming our adapted test is preferable in all situations, it deserves consideration in settings where the treatment effect is anticipated to be primarily cause-specific on the grounds of a well protected type I error for the null of no treatment effect on either hazard, combined with a possibly substantial gain in power. Patients as well as the industry then stand to gain from a speedier trial conclusion.

Under our assumptions, the adapted test is reliable when using diagnostic tests with well-known sensitivity and specificity. The loss of power due to absolute misspecification of the misclassification probabilities in the range of 10 to 20% is comparable to the conservativeness following from using (5.1). This robustness allows one to base p_0 and p_1 estimates on literature data. Alternatively one can estimate these probabilities from a representative diagnostic validation study. Even then, analyses can best be followed by a sensitivity analysis using a predefined probable range for the misclassification rates. Misspecification of the misclassification rates occurring at the design stage can lead to both under- and overestimation of the power, but

the power loss due to a misspecification of up to 20% will stay small in certain settings. Prudence requires one considers the sensitivity issue seriously when deciding on whether or not to use our method. Section 5 of the web-based supplementary material presents an elaborate discussion of sensitivity issues (http://www.biostatistics.oxfordjournals.org). Note that as a rule, misclassification problems can be reduced by aggregating closely related causes-of-death.

Even though the central quantities (4.1) and (4.3) do not rely on simplifying assumptions regarding $\xi(t)$, $p_0(t)$ and $p_1(t)$, the remainder of the text does rely on them to make an adapted design more practical. In practice, more freedom is possible at the analysis stage and one must balance biological relevance with analysis complexity to decide if and which simplifying assumptions are useful. Piecewise constant models for anticipated variations in ξ, p_0 and p_1 imbed much more flexibility and retain the appealing weighted logrank form. Details of such an approach are presented in the supplementary materials, along with a nonparametric estimator for $\xi(t)$ and the incorporation of missing causes-of-death.

A referee has noted that modelling competing risks by means of the cumulative incidence function (Fine and Gray, 1999) enjoys an increasing popularity. We agree it would be interesting to adapt the analog of the logrank test in this setting in a way similar to ours.

While this article has focused on the testing problem, the methods presented can be further developed for estimation of effect sizes in a Cox-type model with specific covariates acting on each of the considered cause-specific hazards (similar to Goetghebeur and Ryan 1995). Obtaining the estimates would however be more difficult and the method would best be supported by software implementation to be practically useful.

In conclusion, we hope this adapted test and corresponding design offer

a cost-efficient alternative for some important classes of problems.

Funding

This work was supported by the IAP research network grant P6/03 of the Belgian government (Belgian Science Policy) [BVR and EG], BVR was also supported through a research fellowship from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen).

Acknowledgments

We are grateful to dr. Cutts, dr. Zaman and the Gambia Government/MRC Laboratories Joint Ethics Committee for allowing us to use the data. BVR wishes to thank CenStat at Hasselt University for the use of their accommodations.

List of Figures

List of Tables

References

ANKER, M. (1997). The Effect of Misclassification Error on Reported Cause-Specific Mortality Fractions from Verbal Autopsy. *International Journal of Epidemiology* **26,** 1090–1096.

ARCHER, L.E. AND RYAN, L.M. (1989). Accounting for Misclassification in the Cause-of-Death Test for Carcinogenicity. *Journal of the American Statistical Association* **84**: 787–791.

CHANDRAMOHAN, D., SOLEMAN, N., SHIBUYA, K. AND PORTER, J. (2005). Ethical issues in the application of verbal autopsies in mortality surveillance systems. *Tropical Medicine and International Health* **10,** 1087–1089.

CUTTS, F.T., ZAMAN, S.M.A., ENWERE, G., JAFFAR, S., LEVINE, O.S., OKOKO, J.B., OLUWALANA, C., VAUGHAN, A., OBARO, S.K., LEACH, A., MCADAM, K.P., BINEY, E., SAAKA, M., ONWUCHEKWA, U., YALLOP, F., PIERCE, N.F., GREENWOOD, B.M., ADEGBOLA, R.A., for the Gambian Pneumococcal Vaccine Trial Group (2005). Efficacy of nine-valent pneumococcal conjugate vaccine against pneumonia and invasive pneumococcal disease in The Gambia: randomised, double-blind, placebo-controlled trial. *The Lancet* **365,** 1139–1146.

CUZICK, D. (2008). Primary endpoints for randomised trials of cancer therapy. *The Lancet* **371,** 2156–2157.

DE FRANCISCO, A., HALL, A.J., ARMSTRONG-SCHELLENBERG, J.R.M., GREENWOOD, A.M. AND GREENWOOD, B.M. (2005). The pattern of infant and childhood mortality in Upper River Division, The Gambia. *Annals of Tropical Paediatrics* **13**: 345-352.

DEWANJI, A. (1992). A Note on a Test for Competing Risks with Missing Failure Type. *Biometrika* **79,** 855–857

DEWANJI, A. AND SENGUPTA, D. (2003). Estimation of Competing Risks with General Missing Pattern in Failure Types. *Biometrics* **59,** 1063-1070.

EBRAHIMI, N. (1996). The effects of misclassification of the actual cause of death in competing risks analysis. *Statistics in Medicine* **15,** 1557–1566.

FINE, J.P. AND GRAY, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94,** 496 -509.

FLEMING, T.R. AND HARRINGTON, D.P. (1991). *Counting processes and survival analysis*. John Wiley & Sons, Inc., New York

GOETGHEBEUR, E. AND RYAN, L. (1990). A Modified Log Rank Test for Competing Risks with Missing Failure Type. *Biometrika* **77,** 207–211.

GOETGHEBEUR, E. AND RYAN, L. (1995). Analysis of Competing Risks Survival Data When Some Failure Types are Missing. *Biometrika* **82,** 821–833.

JAFFAR, S., LEACH, A., GREENWOOD, A.M., JEPSON, A., MULLER, O., OTA, M.O.C., BOJANG, K., OBARO, S. AND GREENWOOD, B.M. (1997). Changes in the pattern of infant and childhood mortality in Upper River Division, The Gambia, from 1989 to 1993. *Tropical Medicine and International Health* **2,** 28–37.

JAFFAR, S., LEACH, A., SMITH, P.G., CUTTS, F. AND GREENWOOD, B. (2003). Effects of misclassification of causes of death on the power of a trial to assess the efficacy of a pneumococcal conjugate vaccine in The Gambia. *International Journal of Epidemiology* **32,** 430-436.

LADOUCEUR, M., RAHME, E., PINEAU, C.A. AND JOSEPH, L. (2007). Robustness of Prevalence Estimates Derived from Misclassified Data from Administrative Databases. *Biometrics* **63,** 272-279.

LATOUCHE, A. AND PORCHER, R.(2007). Sample size calculations in the presence of competing risks. *Statistics in Medicine* **26,** 5370-5380.

LU, K. AND TSIATIS, A. (2005). Comparison Between Two Partial Likelihood Approaches for the Competing Risks Model with Missing Cause of Failure. *Lifetime Data Analysis* **11,** 29-40.

MAUDE, G.H AND ROSS, D.A. (1997). The Effect of Different Sensitivity, Specificity and Cause-Specific Mortality Fractions on the Estimation of Differences in Cause-Specific Mortality Rates in Children from Studies Using Verbal Autopsies. *International Journal of Epidemiology* **26,** 1097–1106.

O'DEMPSEY, T.J., MCARDLE, T.F., LLOYD-EVANS, N., BALDEH, I., LAWRENCE, B.E., SECKA, O. AND GREENWOOD, B. (1996). Pneumococcal disease among children in a rural area of west Africa, The Gambia, from 1989 to 1993. *The Pediatric Infectious Disease Journal* **15,** 431–437.

ROWE, A.K. (2006). Analysis of deaths with an unknown cause in epidemiologic analyses of mortality burden. *Tropical Medicine and International Health* **11,** 540–550.

SCHOENFELD, D. (1981). The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions. *Biometrika* **68,** 316–319.

SCHULGEN, G., OLSCHEWSKI, M., KRANE, V., WANNER, C., RUF, G. AND SCHUMACHER, M. (2005). Sample sizes for clinical trials with

time-to-event endpoints and competing risks. *Contemporary Clinical Trials* **26,** 386–396.

SNOW, R.W., ARMSTRONG, J.R.M., FORSTER, D., WINSTANLEY, M.T., MARSH, V.M., NEWTON, C.R.J.C., WARUIRU, C., MWANGI, I., WINSTANLEY, P.A. AND MARSH, K. (1992). Childhood deaths in Africa: uses and limitations of verbal autopsies. *the Lancet* **340,** 351–355.

SOLEMAN, N., CHANDRAMOHAN, D. AND SHIBUYA, K. (2006). Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization* **84,** 239–245.

SPIRTES P., GLYMOUR C., SCHEINES R., MEEK C., FIENBERG S. AND SLATE E. (1999). *Prediction and Experimental Design with Graphical Causal Models* in Glymour C. and Cooper G.F., eds., *Computation, Causation & Discovery.* AAAI Press/The MIT Press, Menlo Park

WALDEMAR, G., DUBOIS, B., EMRE, M., GEORGES, J., MCKEITH, I.G., ROSSOR, M., SCHELTENS, P., TARISKA, P. and WINBLAD, B. (2007). Recommendations for the diagnosis and management of Alzheimers disease and other disorders associated with dementia: EFNS guideline. *European Journal of Neurology* **14,** e1-e26.

WORLD HEALTH ORGANIZATION (1999). Pneumococcal vaccines - WHO position paper. *Weekly Epidemiological Record* **74,** 177–184, available at
`http://www.who.int/docstore/wer/pdf/1999/wer7423.pdf`.